Introduction to Biostatistics

Biostatistics
-the application of the mathematical tools used in statistics to the fields of biological sciences & medicine.
-Concerned with collection, organization, summarization, & analysis of data.
-Concerned with the interpretation of the data & the communication of information about the data.
-Important for every researcher.
-Seeks to draw inferences about a body of data (population) when only a part of the data (sample) is observed.

Data are numbers which can be measurements or can be obtained by counting, & they represent observations of individuals in our sample
The first step of processing data is called descriptive statistics.

•Population: is the collection or set of all of the values that a variable may have. The entire category under consideration
•Sample: is a part of a population. The portion of the population that is available, or to be made available, for analysis.
(All the subjects in the sample must belong to the population of interest)
•**Representativeness**: the key characteristic of the sample, it is close to the population
(the sample must be representative to the population)
•Sampling: the process of selecting portion of the population (selection of a number of study units/subjects from a defined population)
Sampling bias: excluding any subject without any scientific rational. Or not based on the major inclusion & exclusion criteria.
(If we failed to pick a sample that's representative to the population, then we've committed a mistake called sampling bias)

Questions to Consider (considerations):
-Reference population - to whom are the results going to be applied? (To which population you'll generalize your results)
-Study population - What is the group of people from which we want to draw a sample?
-Sample Size - How many people do we need in our sample? (Small enough to be manageable but large enough to be representative)
-Sampling Method - How will these people be selected?

•Element: The single member of the population (used interchangeably)
•**Sampling frame**: is the listing of all elements/units of the study population
-Sampling depends on the sampling frame.

Types of Sampling Methods:

1.Probability Sampling Methods:
-Random selection process to select a sample from members/elements of the study populations.
-All units of the study population should have an equal or at least a known chance of being selected.
-Requires a sampling frame (Listing all study units)

2.Nonprobability Sampling Methods
-The sample elements are chosen by nonrandom methods.
-More likely to produce a biased sample.
-This restricts the generalization of the study findings.
-Most frequent reasons for use involve convenience & the desire to use available subjects.
-Sometimes we can't or we don't guarantee an equal chance to everybody

Types of **Probability Sampling Methods**

a. **Simple Random Sampling**
•Simplest of probability sampling
•Make a numbered list of all units in the population (sampling frame)
•Decide on the sample size
•Select the required number of sampling units using the lottery method or a random number table
(May not be representative)

b. **Systematic sampling**
•Individuals are chosen at regular intervals from the sampling frame
•Ideally we randomly select a number to tell us the starting point
-every 10th women
(may be not representative)

$$\text{Sampling fraction} = \frac{\text{Sample size}}{\text{Study population}}$$

$$\text{Interval size} = \frac{\text{Study population}}{\text{Sample size}}$$

c. **Stratified sampling**
•If we have study units with different characteristics, then we divide the sampling frame into strata according to these characteristics
•The proportions of individuals with certain characteristics in the sample equal to those in the whole study population
•Random or systematic samples of predetermined sample size will be obtained from each stratum based on a sampling fraction for each stratum
•The different characteristics in the population are represented.

d. **Cluster sampling**
•Selection of study units (clusters) (groups) instead of the selection of individuals (select random clusters)
•All subjects/units in the cluster who meet the criteria will be sampled.
(Clusters are often geographic units)
•Usually used in interventional studies
•Advantages: Sampling frame is not required in this case, Sampling study population scattered over a large area

e. **Multistage sampling**
•More than one sampling method (carried out in phases)
•Does not require an initial sampling frame of the whole population
•Need to know SAMPLING FRAME OF CLUSTERS e.g. province
•Requires sampling frames of final clusters
•Applicable to community based studies (e.g. interviewing people from different villages selected from different areas)

Types of **Nonprobability Sampling Methods**

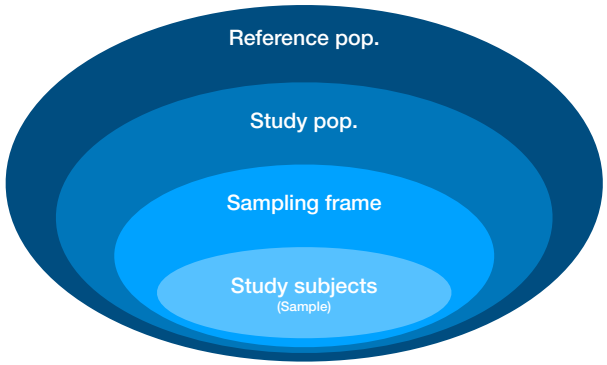
a. **Convenience sampling**
(Accidental or incidental sampling):
•People may or may not be typical of the population, no accurate way to determine their representativeness
•Most frequently used in health research
•Advantages: Saves time & money
•Disadvantage: selection bias + may be not representative

b. **Snowball sampling**
•a method by which the study subjects assist in obtaining other potential subjects (networking)
•Useful in topics of research where the subjects are reluctant to make their identity known, Drug users, Aids patients, etc.
•Useful in rare diseases

c. **Quota sampling**
•The sample is selected by convenience (e.g. the first 50% of males & 50% of females) (you purposefully assign certain percentages because you're afraid a certain portion of the population will not be presented in the sample)
•A mean for securing potential subjects from these strata.
•Variables of interest to the researcher (include subject attributes), such as age, gender, educational background are included in the sample

d. **Purposive sampling** (handpicking, judgmental):
•Subjects are chosen because they are typical or representative of the accessible population, or because they are experts (more knowledgeable) in the field of research topic.
•Qualitative researchers use Purposive sampling
•You purposefully ignore parts of the population
•Doesn't guarantee an equal chance for everybody in the population

The difference between stratified & quota:
In stratified sampling, we respect the actual percentages in the population, but in quota, we purposefully manipulate the percentages



Variable: is an object, characteristic, or property that can have different values. (To be able to study these characteristics or properties in biostatistics, we have to make them numbers, by assigning certain numbers to every characteristic or property)
-A quantitative variable can be measured in some ways. (e.g. how many siblings do you have?)
-A qualitative variable is characterized by its inability to be measured but it can be sorted into categories. (e.g. where are you from? → here we assign a number to every governorate, for example Irbid=1, Karak=2, Amman=3, Ajloun=4 ...)

Types of variables:
1.**Independent variable (IV)** —the presumed cause (of a dependent variable)
2.**Dependent variable (DV)** —the presumed effect (of an independent variable) (depends of the IV)
e.g. smoking (IV) → lung cancer (DV)

Levels of measurement
(in order from weakest/lowest to strongest/highest)

1.Nominal (Discrete/Categorical)
-They are symbols that have no quantitative value, no mathematical meaning.
-Dichotomous: nominal variable that has **2** categories (female vs. male)
-Appropriate statistics: mode, frequency, but we cannot use average
-we don't accept fractions as a result
-E.g. race, blood type, religion, marital status

3.Interval (Continuous)
-They are real numbers & the difference between the ranks can be specified.
-Equal intervals, but no "true" zero.
-attribute can be ordered & the distance between score values is meaningful
-Appropriate statistics: mode, frequency, the median, & the mean
-we accept fractions as a result
-E.g. body temperature

2.Ordinal (Discrete/Categorical)
-attributes can be ordered, numbers have a mathematical meaning.
-The exact differences between the ranks cannot be specified (indicates order rather than exact quantity)
-Appropriate statistics: mode, frequency, & the median; but not the mean.
-we don't accept fractions as a result
-E.g. anxiety level: mild, moderate, severe. (mild=1, moderate=2, severe=3)

4.Ratio (Continuous) (absolute zero)
-Data can categorized, ranked, difference between ranks can be specified & a true or natural zero point can be identified.
-zero point: means that there is a total absence of the quantity being measured.
-we accept fractions as a result
-E.g. scales like weight, total amount of money

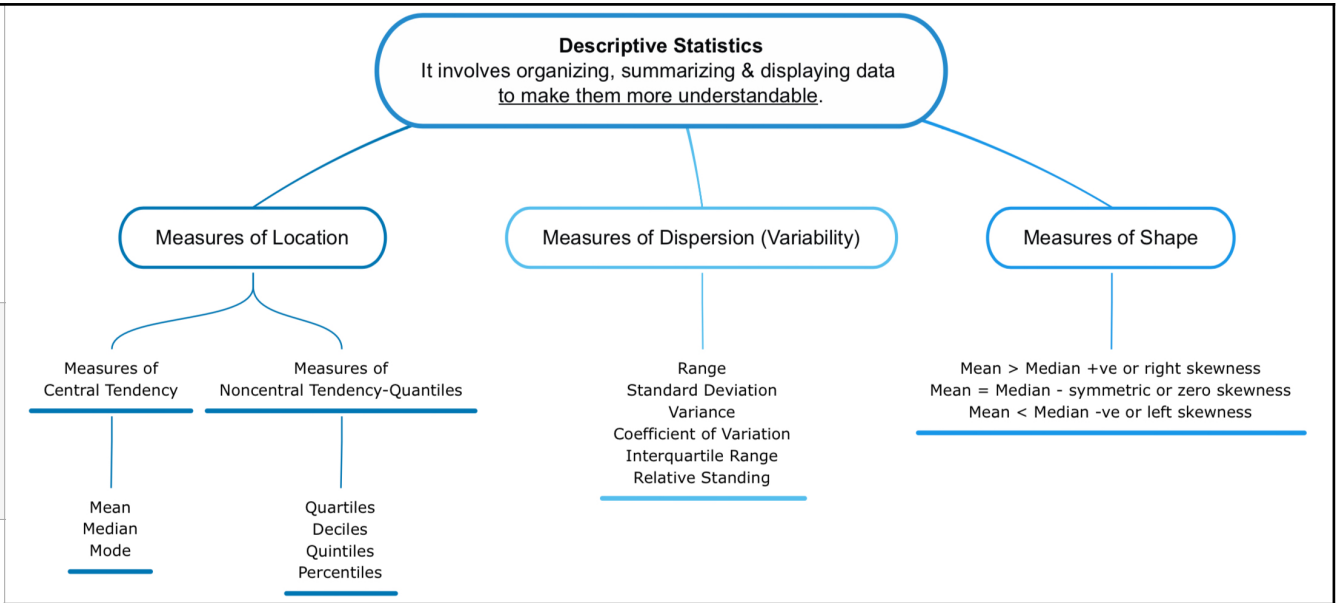
The goal of the researcher is to use the highest level of measurement possible.

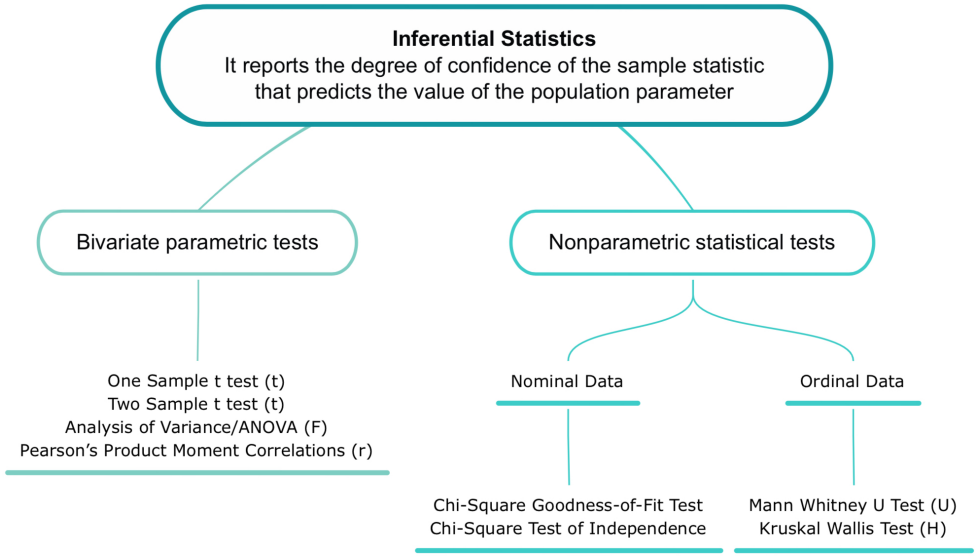
Parameter: is a descriptive measure computed from the data of the **population**.
population mean: μ
population standard deviation: σ
-to determine the population parameters, you take a census of the entire population (very costly)
-the parameters are considered unknown constants

Statistic: is a descriptive measure computed from the data of the **sample**.
sample mean: $\bar{x}$
sample standard deviation: s
-used to estimate the population parameters.

Statistics: a branch of applied mathematics that deals with collecting, organizing, & interpreting data using well-defined procedures in order to make decisions.
Types of Statistics:
- **Descriptive Statistics**
- **Inferential Statistics**

Purposes of statistics
-To describe & summarize information thereby reducing it to smaller, more meaningful sets of data
-To make predictions or to generalize about occurrences based on observations
-To identify associations, relationships or differences between the sets of observations





(Inferential statistics are to make a conclusion about the population using what you have learned from the descriptive statistics from the sample)
Inferential statistics are used to test hypotheses (prediction) about relationship between variables in the population. A relationship is a bond or association between variables.

Research Hypothesis (Alternative) (Ha) (H1)
-A short statement that states a prediction.
-Must always involve at least two variables (IV & DV)
-Must suggest a prediction or explanation of the relationship between the independent variable & the dependent variable.
-It's a translation of the research question.
-Must contain terms that indicate a relationship (e.g., more than, different from, associated with).
-In qualitative research, there is **NO** hypothesis.

Null Hypothesis (H0)
-The opposite of the Ha, it's always a negative statement
-We always try to reject the H0 (this the goal)

Hypotheses Criteria
-Written in a declarative form.
-Written in present tense.
-Contain the population.
-Contain variables (at least one IV & one DV)
-Reflects problem statement or purpose statement.
-Empirically testable.

Hypothesis testing
-A hypothesis is made about the value of a parameter, but the only facts available to estimate the true parameter are those provided by the sample.
-How do we decide whether to reject the H0 or accept it? Through the the inferential statistics
-H0: contains the hypothesized parameter value which will be compared with the sample value.
-H1: will be "accepted" only if H0 is rejected

Two types of **errors** may occur:
α (alpha) (Type I): if you reject H0 when it is true. (False positive)
α is the acceptable margin of type 1 error in your conclusion.
(If you accept a confidence level of 95%, your α is 5%)
β (beta) (Type II): if you accept H0 when it is false. (False negative)
β is the acceptable margin of type 2 error
(usually β is 20% & the power is 80%)

-If you reject the H0 when it's false, no errors
-If you accept the H0 when it's true, no errors

Steps of hypothesis testing: (the doctor only talked about step 6)
1.Formulate H0 and H1
2.Specify the level of significance (α) to be used
3.Select the test statistic
4.Establish the critical value or values of the test statistic needed to reject H0
5.Determine the actual value (computed value) of the test statistic.
6. Make a decision: **Reject H0 or Do Not Reject H0.**

Parametric assumption:
-Dependent variable should be continuous (I/R)
-The independent variable is categorical with two or more levels.
-The observations must be independent
-The observations must be drawn from normally distributed populations
-These populations must have the same variances. Equal variance (homogeneity of variance)
-The groups should be randomly drawn from normally distributed & independent populations
-Distribution for the two or more independent variables is normal.

Advantages of parametric tests
-They are more powerful & more flexible than nonparametric techniques.
-They not only allow the researcher to study the effect of many independent variables on the dependent variable, but they also make it possible to study their interaction.

When the assumptions of the parametric are not achieved, we use the nonparametric

Nonparametric methods
-require no assumptions
-less flexible & less powerful
-called distribution-free methods.
-can be used with small samples

Nonparametric methods must satisfy at least one of the following conditions:
-can be used with nominal data.
-can be used with ordinal data.
-can be used with interval or ratio data when no assumption can be made about the population probability distribution (in small samples).

-There is at least one nonparametric test equivalent to each parametric test
-usually parametric variables compare means & nonparametric variables compare proportions
-Non-parametric tests based on ranks of the data (Work well for ordinal data)

Summary
-Parametric Statistics are statistical techniques based on assumptions about the population from which the sample data are collected.
-Nonparametric Statistics are based on fewer assumptions about the population & the parameters.

Table of statistical tests
(Important)

Level of Measurement (DV)	Sample Characteristics (IV)			
	2 Sample		K Sample (i.e., >2)	
	Independent	Dependent	Independent	Dependent
Categorical or Nominal	X^2	Macnarmar's X^2	X^2	Cochran's Q
Rank or Ordinal	Mann Whitney U	Wilcoxin Matched Pairs	Kruskal Wallis H	Friendman's ANOVA
Parametric (Interval & Ratio)	t test between groups	t test within groups	1 way ANOVA between groups	1 way ANOVA (within or repeated measure)

Raneem Al-Zoubi